Language Recognition

说话人识别与语种识别是相似的任务，所以说话人识别的许多方法都能用在语种识别中。

特征

BNF : bottleneck feature.

提取自ASR神经网络的隐藏层的输出

Acoustic Feature : MFCC, Fbank, etc.

Tandem feature :

BNF + acoustic feature

模型

TDNN

LSTM/BLSTM

CNN

打分/分类

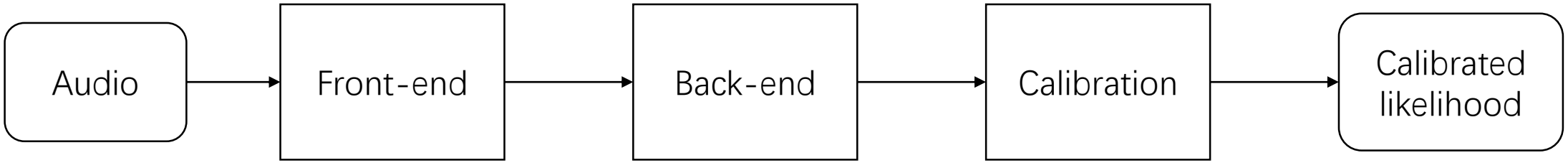
PLDA

LDA

SVM

NN

基本步驟



Front-end

- VAD (SAD, Speech Activity Detection)
- Feature Extractor

Back-end

- WGB (Weighted Gaussian)
- LDA/PLDA
- SVM
- NN

Calibration

- Multiclass logistic regression

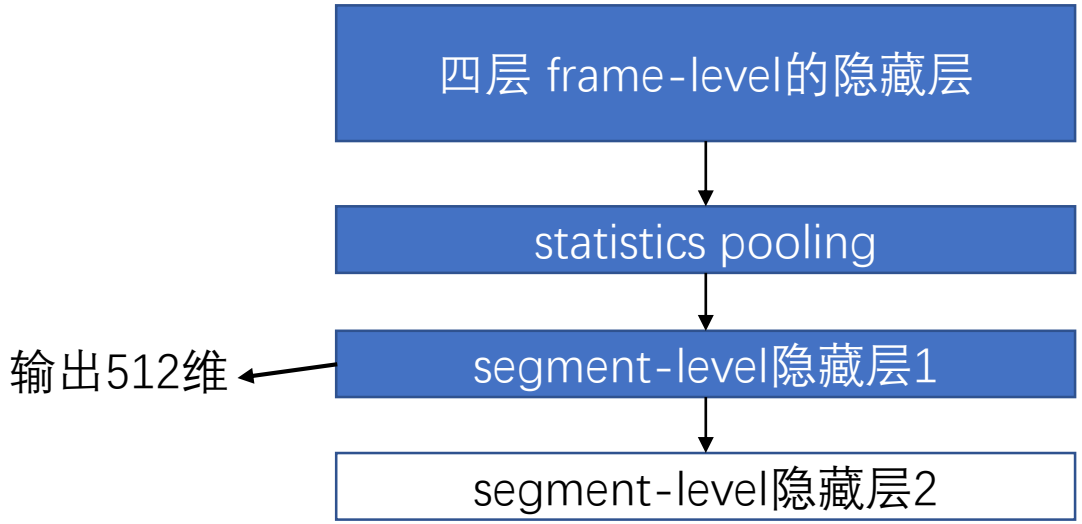
Triplet Neural Networks

https://www.isca-speech.org/archive/Interspeech_2019/abstracts/2437.html

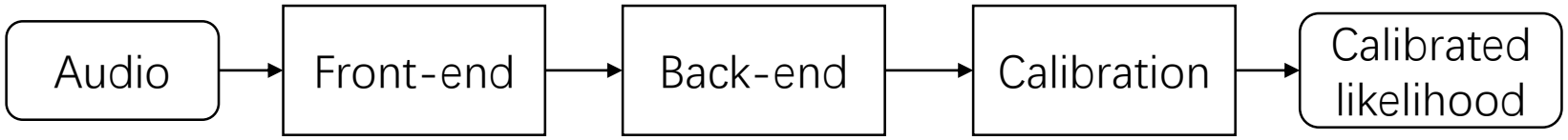
Triplet Neural Networks

Front-end

- VAD (SAD, Speech Activity Detection)
- Senone DNN Bottleneck Extractor
  (输出80维)
- Language embeddings extractor
  用BNF训练的，能够区分49种语言的网络



Triplet Neural Networks

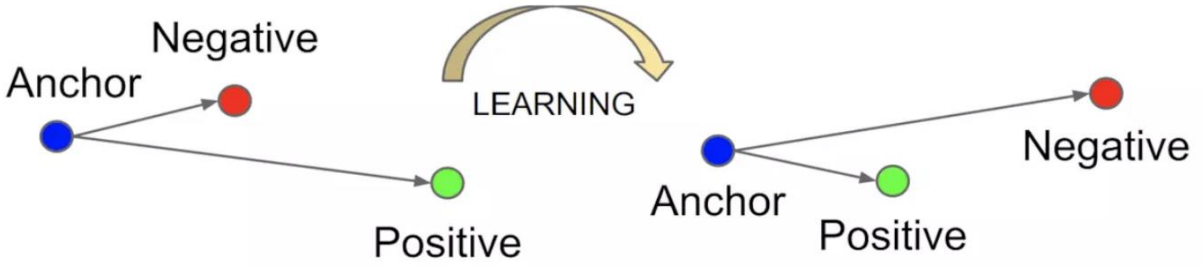


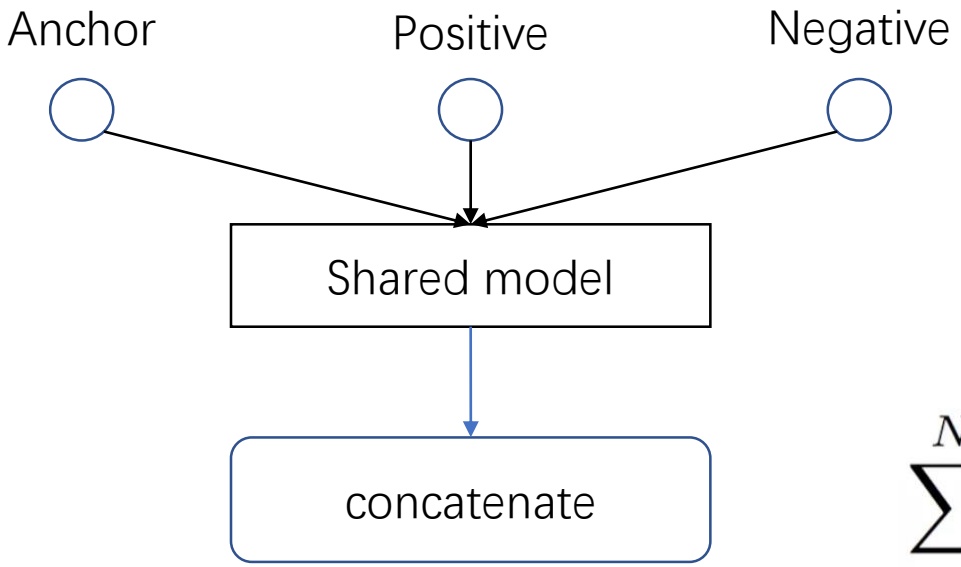
Back-end

- LDA将512维降到48
- Normalization
- Triplet NN

Triplet loss

取一个数据，为Anchor
再取一个同类别的数据，为Positive
再取一个不同类别的数据，为Negative



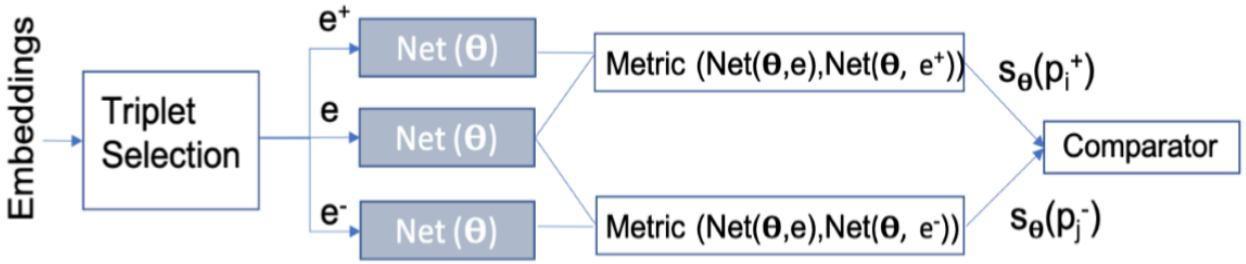


$$\sum_{i}^{N}\left[\|f(x_i^a)-f(x_i^p)\|_2^2-\|f(x_i^a)-f(x_i^n)\|_2^2+\alpha\right]_+$$

Triplet loss

取一个数据，为e
再取一个同类别的数据，为e^+
再取一个不同类别的数据，为e^-

网络参数： θ
p_i^+ ： $pair(e,e^+)$
$S_\theta(p_i^+)$: p_i^+ 的相似性测度
m^+： 正样本对数量



Shared model

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}} \frac{1}{m^+ m^-} \sum_i^{m^+} \sum_j^{m^-} \sigma(\alpha(s_\Theta(p_i^+) - s_\Theta(p_j^-)))$$

Triplet Neural Networks

Back-end	Results (Cllr / EER%)			
	LRE09	LRE15	LRE15-nofre	LRE17
PLDA	0.135/3.07	**0.231**/**5.91**	0.188/**4.60**	0.285/7.35
DNN	0.149/3.38	0.345/8.83	0.258/7.36	0.359/8.11
TripleNet – Rand	0.129/2.72	0.443/8.11	0.351/7.18	0.438/8.02
TripleNet – HM1	0.119/**2.29**	0.372/6.33	0.285/5.49	0.351/6.73
TripleNet – HM2	**0.112**/2.61	0.274/6.36	**0.183**/4.95	**0.283**/**6.72**

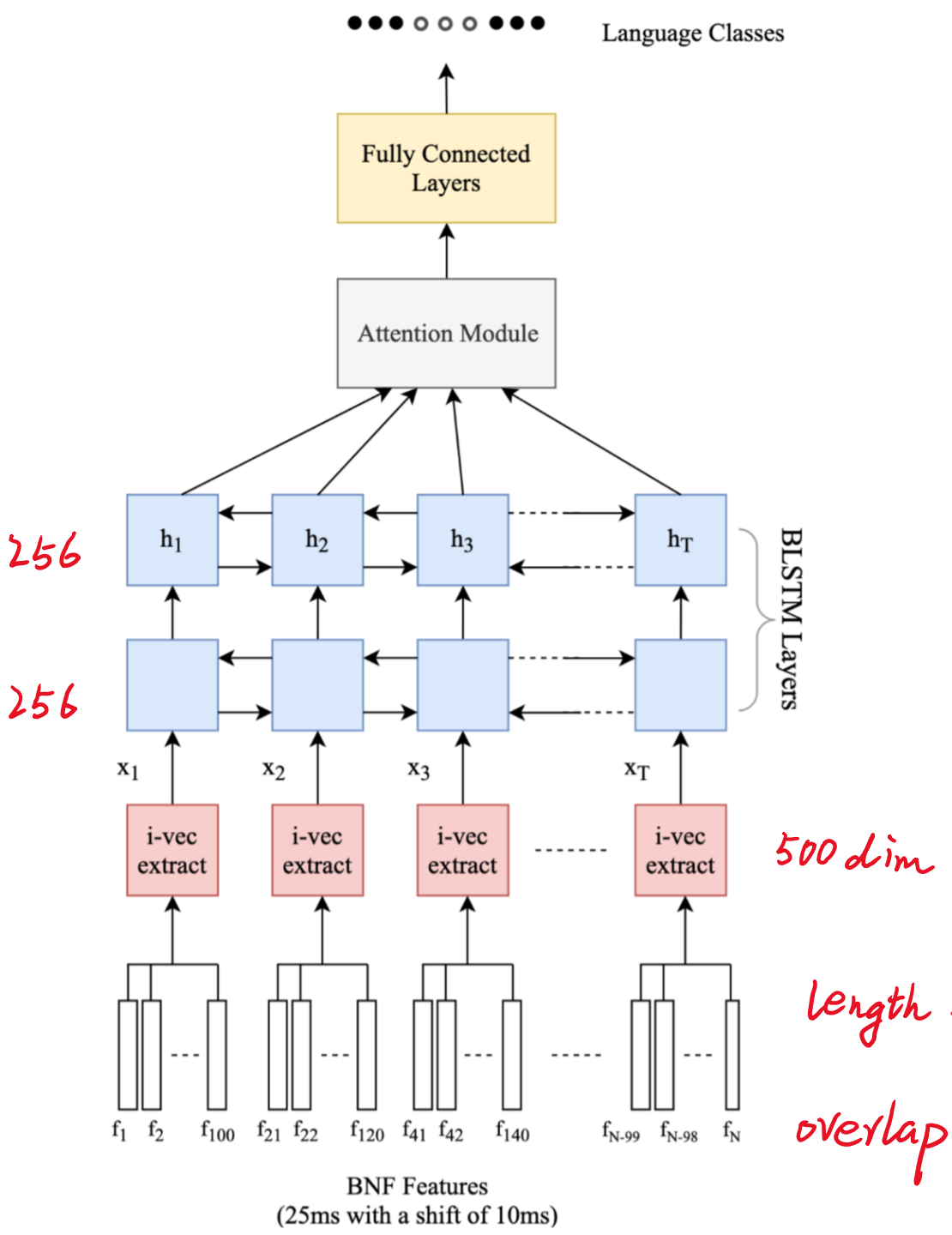
Rand：一组选***P***种语言，一共***G***组。若数据中一共有***N***种语言，***N = P·G***

Hard Negative Mining Selection

1. 用每***P***个语言中的***K***个样本来做triplet
2. 和1相似，但从数据中找出更难以区分的样本做triplet

Attention Based Hybrid i-Vector BLSTM

https://www.isca-speech.org/archive/Interspeech_2019/abstracts/2371.html



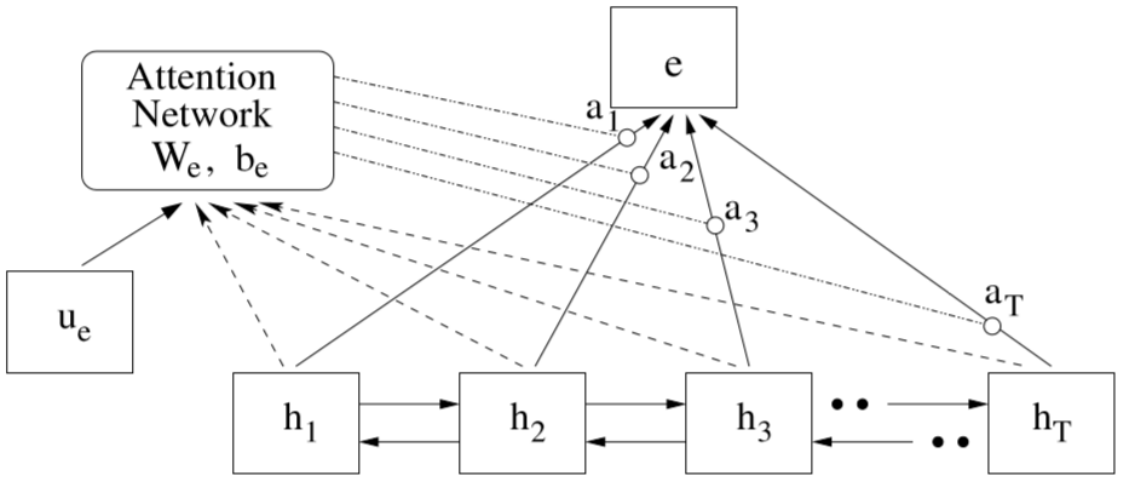
BNF:

80维

来自于为ASR训练的DNN

该模型使用的特征是MFCC

Attention Module



$$\mathbf{u}_t = tanh(\mathbf{W}_e\mathbf{h}_t + \mathbf{b}_e)$$

$$a_t = \frac{exp(\mathbf{u}_t^T\mathbf{u}_e)}{\sum_t exp(\mathbf{u}_t^T\mathbf{u}_e)}$$

$$\mathbf{e} = \sum_t a_t\mathbf{h}_t$$

e相当于h_t的加权平均

a_t为序列添加了注意力范围

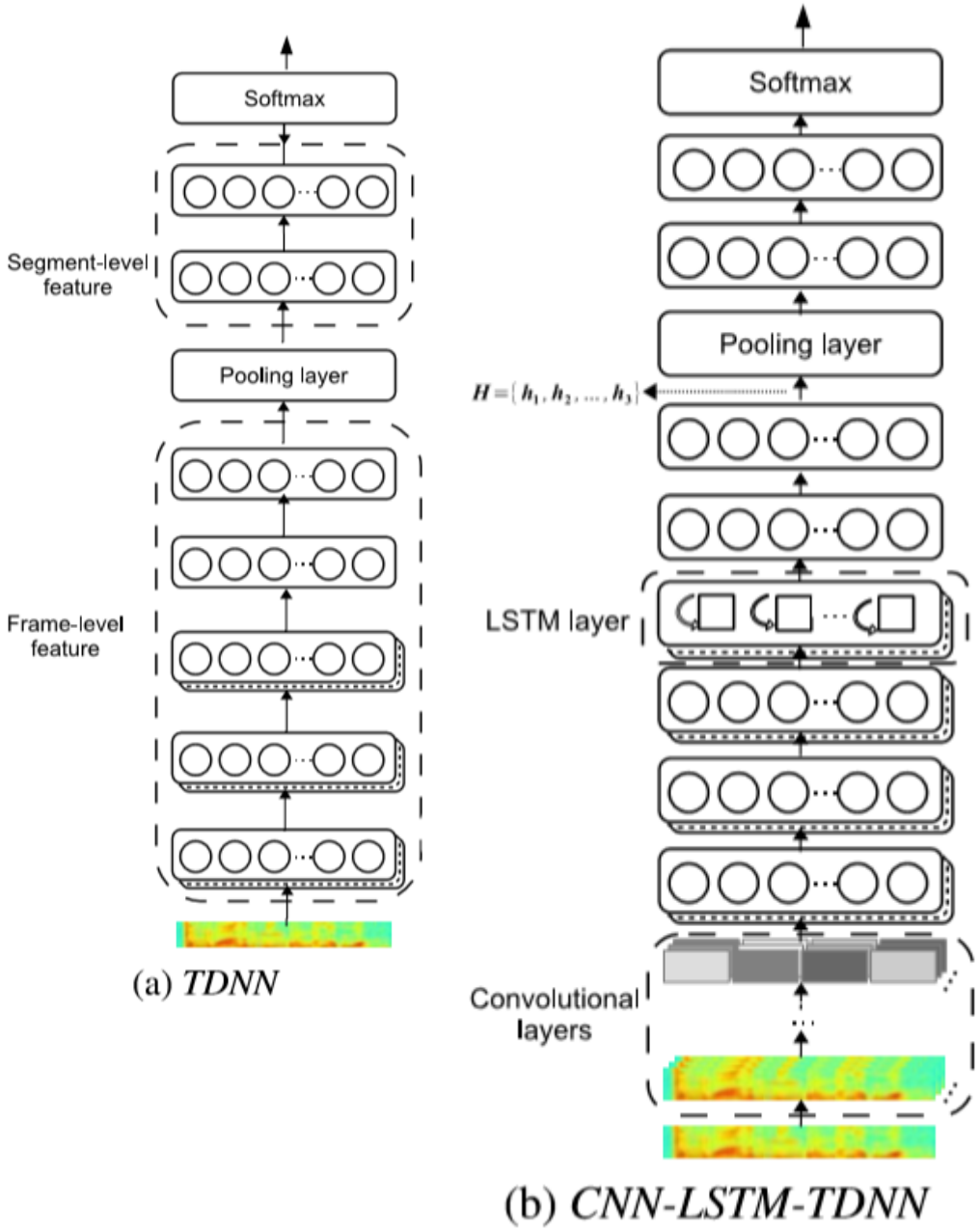
Dur./Model	LDA-SVM [21]	LSTM [28]	i-BLSTM
Accuracy (%)			
3	53.84	54.74	**54.80**
10	72.36	72.58	**75.89**
30	**82.98**	76.10	82.27
1000	**56.23**	42.86	54.07
overall	67.86	64.74	**68.65**
C_{avg}			
3	0.53	0.55	**0.50**
10	0.27	0.35	**0.26**
30	**0.13**	0.28	0.18
1000	0.54	0.79	**0.50**
overall	0.37	0.48	**0.36**
EER (%)			
3	**13.40**	15.39	15.47
10	6.47	8.70	**6.32**
30	**3.50**	7.25	3.67
1000	15.35	26.27	**14.71**
overall	**9.26**	14.38	9.65

SNR/Model	LDA-SVM [21]	LSTM [28]	i-BLSTM
No noise	72.36	72.1	**75.89**
Partially Noisy			
5dB	53.31	56.50	**59.79**
10dB	55.76	60.42	**63.02**
15dB	58.49	62.61	**65.90**
20dB	59.78	64.61	**68.16**
overall	56.83	61.03	**64.22**
Noisy			
5dB	47.93	48.36	**51.58**
10dB	53.77	56.30	**59.86**
15dB	57.82	61.63	**64.30**
20dB	60.00	65.28	**67.72**
overall	54.88	57.89	**60.87**

Time-frequency attention mechanism for TDNN and CNN-LSTM-TDNN

https://www.isca-speech.org/archive/Interspeech_2019/abstracts/1256.html

CLSTM x-vector



CNN通过反向传播学习到时序特征

Time Attention

$\boldsymbol{H}$: input
$\boldsymbol{T_A}$: 注意力系数矩阵
$\boldsymbol{g()}$: 激活函数，ReLU

$$\mathbf{T}_A = softmax(g(\boldsymbol{H}^T \boldsymbol{W}_1^t)\boldsymbol{W}_2^t)$$

$$\mathbf{E}_t = \boldsymbol{H}\boldsymbol{T}_A$$

$\boldsymbol{W_1^t}$:$d_h \times d_a$
$\boldsymbol{W_2^t}$:$d_h \times 1$

Frequency Attention

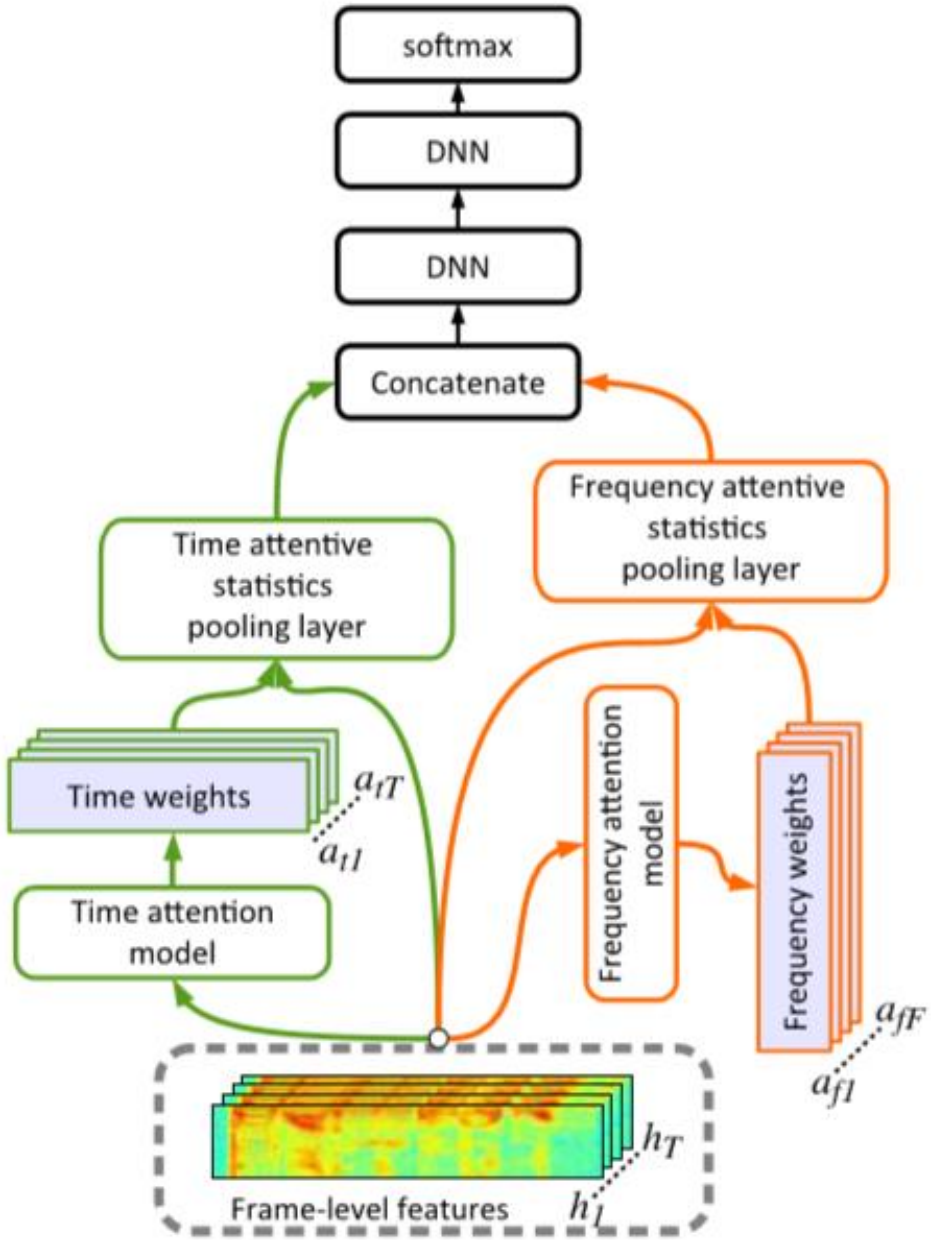
$$\mathbf{F}_A = softmax(g(\boldsymbol{H}^T \boldsymbol{W}_1^f)\boldsymbol{W}_2^f)$$

$\boldsymbol{W_1^f}$:$d_h \times d_a$
$\boldsymbol{W_2^f}$:$d_a \times d_f$

E.g.

$$d_h = 1500, d_f = 2$$

两个频率带：[1:750], [751:1500]
令$F_A = [a_{f1}, a_{f2}]$
$a_{f1} \times h \in [1:750], a_{f2} \times h \in [751:1500]$
h是weight，用来计算期望和方差



***s*()代表打分函数**

$$s(u) = (1-\alpha)s^f(u) + \alpha s^t(u)$$

Table 1: *Performance results for time and frequency attention.*

	3s			10s			30s		
System	**ER**	**C_avg**	**EER**	**ER**	**C_avg**	**EER**	**ER**	**C_avg**	**EER**
GMM i-vector	43.33	18.49	16.12	19.32	9.31	7.04	8.83	4.18	2.50
DNN i-vector	31.42	13.46	10.75	9.82	4.28	3.24	4.08	1.39	1.11
DBF DNN i-vector	23.73	9.35	7.73	8.06	3.17	4.94	2.46	1.11	0.78
DNN x-vector	25.90	10.31	9.03	11.17	3.56	3.38	5.79	1.75	1.71
DNN x-vector time	24.84	9.96	9.12	11.03	3.67	3.33	6.67	1.86	1.71
CLSTM	19.51	7.09	6.67	6.02	1.64	1.66	2.46	0.81	0.78
CLSTM time	19.46	7.14	6.76	5.38	1.68	1.76	2.69	0.70	0.83
CLSTM fre2D	19.69	7.32	6.90	5.05	1.64	1.71	1.90	0.49	0.64
CLSTM fre8D	19.51	6.60	6.58	5.14	1.48	1.66	1.90	0.53	0.69
CLSTM fre16D	19.69	6.84	6.67	5.05	1.39	1.62	2.13	0.53	0.60
CLSTM fre23D	19.14	6.47	6.48	4.91	1.40	1.48	1.71	0.42	0.55
CLSTM fre32D	19.05	6.29	6.16	4.77	1.33	1.43	1.85	0.43	0.55

Language Recognition

END